Regional Relation Modeling for Visual Place Recognition

Yingying Zhu*
Shenzhen University
zhuyy@szu.edu.cn

Biao Li
Shenzhen University
libiaoup@gmail.com

Jiong Wang
Zhejiang University
liubinggunzu@gmail.com

Zhou Zhao
Zhejiang University
zhaozhou@zju.edu.cn

ABSTRACT

In the process of visual perception, humans perceive not only the appearance of objects existing in a place but also their relationships (*e.g.* spatial layout). However, the dominant works on visual place recognition are always based on the assumption that two images depict the same place if they contain enough similar objects, while the relation information is neglected. In this paper, we propose a regional relation module which models the regional relationships and converts the convolutional feature maps to the relational feature maps. We further design a cascaded pooling method to get discriminative relation descriptors by preventing the influence of confusing relations and preserving as much useful information as possible. Extensive experiments on two place recognition benchmarks demonstrate that training with the proposed regional relation module improves the appearance descriptors and the relation descriptors are complementary to appearance descriptors. When these two kinds of descriptors are concatenated together, the resulting combined descriptors outperform the state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies → Visual content-based indexing and retrieval**; **Image representations**.

KEYWORDS

Visual place recognition; Content-based image retrieval; Convolutional neural network; Relation modeling

ACM Reference Format:
Yingying Zhu, Biao Li, Jiong Wang, and Zhou Zhao. 2020. Regional Relation Modeling for Visual Place Recognition. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397271.3401176

*Corresponding author.

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401176

1 INTRODUCTION

In this paper, we tackle the Visual Place Recognition (VPR) task which has consistently attracted considerable attention in computer vision [2, 4, 8, 56] and robotics communities [9, 10, 35, 60]. VPR is traditionally cast as a large-scale image retrieval problem [62] and the goal is to localize query images by searching a large geo-tagged database. Essentially, VPR requires discriminative image embedding to encode the geo-informative objects characterizing a place. In the urban environment, however, buildings with repetitive architecture and the confusing objects such as cars and pedestrians degrade the image representations. The appearance of a place may also change dramatically with the variances of viewpoint, season and illumination condition. Thus VPR is a very challenging task.

Early VPR works [4, 22, 26, 53, 56] are mostly based on the classical image retrieval methods such as bag-of-visual-words model [49] and VLAD descriptors [21] to enable robust place recognition. Arandjelović *et al.* [2] incorporate VLAD coding with differentiable operations in the deep convolutional neural network (CNN) architecture, and the proposed NetVLAD descriptors show powerful representative ability on VPR and image retrieval datasets. The Contextual Re-weighting Networks (CRN) [24] were proposed to alleviate the influence of confusing objects for the NetVLAD descriptors. The attention-based pyramid aggregation network [63] sum-aggregates attentive region features to overcome the influence of repetitive building architecture and confusing objects. Although effective, these works regard the objects in an image as separate and aggregate them in an orderless manner, while the relations between objects are lost. The loss of relations comes mainly from two aspects. First, the local handcrafted features [7, 33] or CNN features are inefficient to model the object relations. Second, the orderless characteristic of their aggregation methods (*e.g.* Sum-aggregation is most commonly) discards spatial relation information.

Essentially, the similarity comparison of global image representations is equivalent to the matching of local features [50, 51], and different aggregation methods (*e.g.* NetVLAD or APANet) define different matching strategies. As shown in Figure 1, two images at different places contain four buildings similar to each other. Traditional image retrieval methods compare the similarity of the objects in two images, and these two images may have considerable similarity and be falsely matched to the same place. But when we take the spatial relation into consideration (*e.g.* building A is close to B and E, while C is far from D and there is no E around it), we can easily recognize they are not in the same place. As is often the case in the urban environment, buildings are similar at different places. Moreover, the appearance of buildings may change dramatically

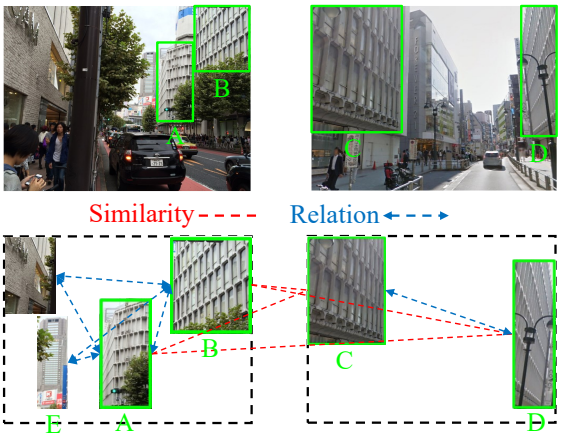


Figure 1: Motivation of our method. The first row shows two images from Tokyo 24/7 dataset at different places but having similar buildings. The second row describes the similarity comparison (red dashed line) of two images and the relation information (blue dashed line) between objects. Two images might get considerable similarity and be falsely matched when only the building appearance is considered. But when the relationships between buildings are considered, the similarity becomes very low.

with times and illumination conditions, while the visual relationships between buildings won't change much under these variances, thus they can serve as robust cues for place recognition.

Our work is based on the regional pooling methods [15, 52, 63] which show competitive performance in image retrieval task and inspired by the relationship modeling framework which is widely adopted in the Visual Relation Detection (VRD) task [11, 34, 59]. We propose a regional relation module to imitate the relationship modeling process in VRD task and implicitly model the regional relation information without any supervision on objects, object pairs or object relationships. As shown in Figure 2, we adopt the spatial grids of pyramid pooling to define regions and model "all-to-all" region-pairs relations by the relation module. The regional relation module generates relation features of all region-pairs and converts the Convolutional Feature Maps ($\mathcal{C}$-FM) to Relational Feature Maps ($\mathcal{R}$-FM), which contain high-level information about object relations. Further on, we design a cascaded pooling method, consisting of K-Max pooling at row-wise and sum pooling, to effectively aggregate the $\mathcal{R}$-FM and get the global relation descriptors.

In summary, our contributions are threefold.

- First, we consider the visual relationship as a robust cue for VPR and propose a regional relation module to model the relation information between regions. The generated relation features can be aggregated by traditional aggregation methods and provide a more reasonable matching assumption than convolutional features for image retrieval task.
- Second, we design a cascaded pooling method to get compact and discriminative relation descriptors. The cascaded pooling method consistently outperforms the commonly used pooling methods for aggregating the R-FM.
- Third, the proposed relation descriptors are complementary to the traditional appearance descriptors. When concatenating them together, the resulting combined descriptors outperform state-of-the-art works on two place recognition benchmarks.

2 RELATED WORK

2.1 Image retrieval-based place recognition

In this paper, we consider the image retrieval-based visual place recognition (VPR) task, which has been widely studied in the computer vision community and is different from some related but not identical works such as the classification-based [17, 46] and 2D-3D matching-based localization task [31, 44]. Traditionally, VPR works adopt generic image retrieval methods such as bag-of-visual-words model [49], VLAD descriptors [21], region features [52], and incorporate the discovering of distinctive or confusing features [4, 22, 26, 63], analysis of repetitive architectures [55, 56] and viewpoint changes [53, 54].

Based on the CNN architecture, NetVLAD descriptors [2] were proposed and end-to-end trained on the Google Street View datasets for efficient place recognition. The Contextual Re-weighting Networks (CRN) [24] were proposed to alleviate the influence of confusing objects for the NetVLAD descriptors. For the same purpose, Attention-based Pyramid Aggregation Networks (APANet) [63] re-weight region features with attention block to improve the discrimination of regional pooling method [14, 52]. However, all these works ignore the relations between objects in an image. Spatial matching [38, 47] is a classical method to verify the spatial configurations of local features for image retrieval. But the recent deep local-feature based methods [36, 47] are usually two-stage and cause additional memory and computation overhead after training the retrieval network. Different from these works, we model regional relations in an end-to-end training architecture. We propose a regional relation module which converts the region features to relation features with high-level information about regions and their relations. The relation features can be aggregated by commonly-used aggregation methods and show excellent performance.

2.2 Visual relational reasoning

Facilitated by the maturity of visual recognition tasks [18, 41], high-level visual relational reasoning tasks such as visual relation detection (VRD) [11, 30, 34] and visual question answering (VQA) [1, 16, 23] are extensively studied recently. The neural network architecture is hard to model relations itself, so the Gated Graph Sequence Neural Networks [29], Interaction Networks [6], Non-local Neural Network [57] and Relation Networks (RN) [19, 43] are proposed to enable relation-centric computation. Among these works, RN is a simple, plug-and-play module for effective relational reasoning and shows super-human performance on VQA task.

The design of our regional relation module is inspired by the VRD works and somewhat similar to RN, and the differences are mainly in two aspects. First, we aim to retrieve the place of an image rather than explicitly predict the relationships, and the relation information is implicitly included in our relation descriptor as a robust cue. Second, RN models the relation of local CNN features while our regional relation module models the regional relation,

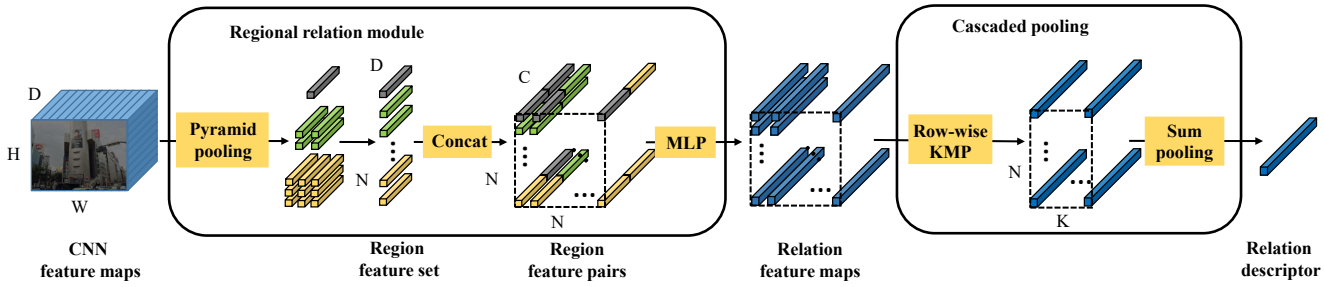


Figure 2: Illustration of our network architecture for regional relation modeling. Regional relation module models regional relations and converts the convolutional feature maps to relational feature maps. The cascaded pooling method is proposed to aggregate the relational feature maps and get a global relation descriptor. Row-wise KMP denotes K-Max pooling at row-wise as described in Section 4.2.

which is computation effective and feasible for images with different resolutions in image retrieval datasets. Compared to RN, we further design a cascaded pooling method to effectively aggregate the relational feature maps.

3 PRELIMINARY

Before we describe the proposed regional relation module and the cascaded pooling method (Section 4), we first briefly introduce the matching assumption of traditional image retrieval methods (Section 3.1) and the commonly-used visual relationship modeling framework which inspires our work (Section 3.2).

3.1 Traditional matching assumption

In this subsection, we focus on the popular aggregation methods [2, 5, 52] based on the deep CNN features and their underlying matching strategies for image retrieval task. The commonly used sum pooling method [5] sum-aggregates local CNN features to a global descriptor as formulated:

$$F_s(I) = \sum_{y=1}^{H}\sum_{x=1}^{W} f_{x,y}, \tag{1}$$

in which H, W are spatial size of the CNN feature maps (C-FM) generated from image I, f is local features on the C-FM. The similarity comparison of two global image descriptors is equal to:

$$\langle F_s(I_1), F_s(I_2)\rangle = \sum_{f_i \in I_1}\sum_{f_j \in I_2} \langle f_i, f_j\rangle, \tag{2}$$

where $\langle , \rangle$ denotes inner product operation. It can be seen that sum pooling provides an "all-to-all" matching kernel, where the similarity comparison of two global descriptors corresponds to the cross-matching of all local features.

Similarly, region-based pooling methods, such as R-MAC [52] and APANet [63], measure the "all-to-all" similarities of region features. The matching strategy of global max pooling is based on the comparison of maximum activation at each channel on the feature maps [50], and VLAD is based on a selective matching kernel as described in [3, 51]. It is intuitive that these aggregation methods endow the aggregated descriptors with invariance to scaling and translation by discarding the spatial information. But the local CNN features or region features themselves are inefficient to model the object relations, therefore the relation information in an image is greatly discarded.

In summary, the matching assumption of existing works can be outlined as two images are matching each other if they contain enough similar objects, no matter how these objects are spatially arranged or what their relationships are. This assumption is distinct to human perception because human equally focus on the relationships between objects in an image. In this paper, we aim to model the relationships of region features and the matching of the proposed relational region features provides a more reasonable matching assumption.

3.2 Visual relationship modeling framework

Here we consider the visual relationship modeling (VRM) framework [34, 59] which models objects and predicates separately. The relationships in an image are <subject, predicate, object> triplets, where the subject, object are objects in the image and predicate describes their interaction, *e.g.* <Man, Hold, Baseball>. Usually the objects in an image are first detected by a detector [41], then the object pairs are constructed for predicate prediction. Supposing an object pair (o_i, o_j) is detected, the region of interest (ROI) features are first concatenated (f_{o_i}, f_{o_j}) and passed to multilayer perceptron (MLP) to get the relation feature. Then the relation feature is sent to a soft-max classifier for predicting the predicate score. The r^{th} predicate score is calculated by:

$$S(i,j,r) = \frac{\exp(w_r^T MLP(f_{o_i}, f_{o_j}))}{\sum_{t=1}^{P}\exp(w_t^T MLP(f_{o_i}, f_{o_j}))}, \tag{3}$$

where w_r^T is the parameter of r^{th} category in the classifier, P is the number of predicate categories.

This framework is widely adopted in VRD, VQA tasks [43] and some self-supervised learning works [12, 37] for relationship prediction, and it also inspires us to model object relations in an image.

4 METHOD

4.1 Regional relation module

4.1.1 Region features. For VPR datasets, there are only image-level supervisions such as GPS coordinates, and no supervisions on objects or relationships in the images are available. The reason is that most of the object categories in PASCAL VOC [13] or the Visual Genome [27] dataset, such as "person" and "bus", are confusing objects for recognizing place. The trained Faster-RCNN is not able to detect the geo-informative buildings. So we define the notion of objects and object pairs in a fully unsupervised way. We first analogize the region features generated by pyramid pooling to the object features in the VRM framework. As illustrated in Figure 2, having $\mathcal{C}$-FM with size of $W \times H \times D$ from CNN's last convolutional layer, we adopt pyramid pooling with n-scale spatial grids to get N region features. Specifically, we perform regional max pooling in the spatial grids and adjacent regions in each spatial grid are 50% overlapping. The size of pooling window is $[\lceil 2 \times W/(n+1) \rceil, \lceil 2 \times H/(n+1) \rceil]$, and the pooling stride is $[\lceil W/(n+1) \rceil, \lceil H/(n+1) \rceil]$, where $\lceil \rceil$ is the ceiling function. In this way, $n \times n$ region features can be obtained. The region feature set $\mathbf{f}_\Omega$ is formulated as follow:

$$\mathbf{f}_\Omega = \{f_{r,1} ... f_{r,s_1^2} ... f_{r,N}\}, \text{ with } N = \sum_{i=1}^{n} s_i^2, \tag{4}$$

where $f_{r,j}$ is the j^{th} region feature generated from the spatial grids, and s_i is the side length of i^{th} scale spatial grid. After that, we send $\mathbf{f}_\Omega$ to the relation module for relation modeling.

4.1.2 Relation module. For regional relation modeling, We make a hypothesis that every two regions have informative relations, and we take the "all-to-all" regional relations into consideration. In this way, we imitate the relation modeling process of VRM framework by a relation module, where the region features are concatenated each other and their relationship are modeled by the MLP. Compared with the CNN feature maps ($\mathcal{C}$-FM), the obtained relational feature maps ($\mathcal{R}$-FM) contain higher-level object appearance information and object relation information. As shown in Figure 2, having the region feature set $\mathbf{f}_\Omega$ with size of $N \times D$, we concatenate these region features each other to get the region feature pairs with size of $N \times N \times C$ ($C = 2 * D$). Then 1×1 convolutional layers (MLP) with shape preserving are applied to generate the relational feature maps ($\mathcal{R}$-FM), in which each relation feature $f_{\mathcal{R}_{i,j}}$ is formulated as:

$$f_{\mathcal{R}_{i,j}} = MLP(f_{r,i}, f_{r,j}). \tag{5}$$

Then we adopt cascaded pooling to aggregate the $\mathcal{R}$-FM and get the global relation descriptor. The difference to VRM framework exists that we aggregate the relation features for a discriminative global descriptor rather than explicitly predict their relationships with a classifier.

4.2 Cascaded pooling

The commonly used Global Max Pooling (GMP) or Global Average Pooling (GAP, the same as sum pooling) can be used to aggregate the $\mathcal{R}$-FM for a global relation descriptor. But when the "all-to-all" regional relations are taken into consideration, the $\mathcal{R}$-FM contain not only rich information but also a lot of noise, so it is inferior to directly adopting GMP or GAP for aggregation. Additionally, considering the spatial characteristic of $\mathcal{R}$-FM, we propose a cascaded pooling method to preserve as much useful information as possible and prevent the influence of confusing relations.

Note that the spatial characteristic of $\mathcal{R}$-FM is quite different from $\mathcal{C}$-FM. As illustrated in Figure 2, the $N \times N$ relation features in $\mathcal{R}$-FM are pseudo-symmetric because the relation features symmetrical about the main diagonal represent relations of the same two regions ((r_i, r_j) or (r_j, r_i)), and the difference is their order. Another spatial characteristic is that each row or column of the $N \times N$ relation features represents the relations of one region to all regions. Considering these properties, we propose a two-step pooling method, the cascaded pooling, which consists of K-Max pooling at row-wise (similar to column-wise) and sum pooling for aggregation. K-Max pooling (KMP) selects K strongest activations in the given spatial size, and max pooling can be seen as a special case of KMP when $K = 1$. In cascaded pooling, KMP at row-wise gets $N \times K$ most strong activations for each channel on the $\mathcal{R}$-FM, and then sum pooling gets a C-dimensional global relation descriptor $F_\mathcal{R}$, which is formulated as:

$$F_\mathcal{R} = \sum_{i=1}^{N} \sum KMP(f_{\mathcal{R}_{i,1}}, f_{\mathcal{R}_{i,2}}, ..., f_{\mathcal{R}_{i,N}}). \tag{6}$$

At last the relation descriptor is ℓ_2-normalized for training or testing. The interpretation of our cascaded pooling is that we first select salient relations for each object (region) and then sum-aggregate all these relations to a global relation descriptor.

4.3 Composite matching assumption

Compared to the regional pooling methods [52, 63] that match the similarity of region features $f_{r,i}$ in two images as described in Section 3.1, the matching of our regional relation features $f_{\mathcal{R}_{i,j}}$ provides a composite matching assumption that two matching images should have similar region pairs and the pair-wise relationships should match accordingly. This new assumption is obviously more reasonable for human common sense.

There are two reasons that we adopt the region features instead of local CNN features to analogize the object features. First, region features avoid the over-counting problem caused by the repetitive architecture of buildings and show competitive performance in image retrieval works [52, 63]. Second, adopting the local CNN features for relation modeling gets a huge $\mathcal{R}$-FM with size of $HW \times HW \times C$, which brings high memory and computation overheads, especially for VPR datasets.

4.4 Training and testing

4.4.1 Training. The proposed regional relation module and cascaded pooling method are composed of differentiable operations. Therefore our architecture is end-to-end trained on the Google Street View training datasets with the weakly supervised triplet loss [2], where the goal is to make the matching images closer and the dis-matching images far from each other in the descriptor space. Triplet loss has shown its effectiveness in many vision tasks such as face identification [45] and image retrieval [15, 32]. Learning to rank the positive and negative images in the triplets enables the neural networks to produce discriminative descriptors.

Given a query image q, we construct a training tuple $(q, p, \{n_j\})$, where p is a positive image matching the query and $\{n_j\}$ is a set of negative images. The positive and negative images are selected according to their GPS coordinate annotations, which provide weak form of supervision. Detailed descriptions about mining the training tuples can be found in [2]. The weakly supervised triplet ranking loss is defined as follow:

$$L_{(q,p,\{n_j\})} = \sum_j max(0, d_\theta^2(q,p) + m - d_\theta^2(q,n_j)), \quad (7)$$

where d_θ^2 denotes the Euclidean distance in the descriptor space, and m is a scalar representing the margin. When any of the $\{n_j\}$ is close to q in the descriptor space, exceeding the margin compared to the positive image, there is a loss to be back-propagated. In this way, the proposed relation descriptor is optimized in an end-to-end manner.

4.4.2 Testing. Since the proposed relation descriptors focus on different cues to traditional appearance descriptors, we found they are complementary to each other. So we concatenate the appearance descriptors with relation descriptors to get the combined descriptors, which have large accuracy improvement. Specifically, we choose the Pyramid Aggregation (PA) descriptors (directly sum-aggregating the region features in Figure 2) as appearance descriptors because they show better performance than the GMP or GAP descriptors and can be obtained together with relation descriptors in a single forward pass. It's worth mentioning that optimizing the combined descriptors end-to-end yields poorer performance, so we directly optimize the relation descriptors. We also found regularizing PA descriptors with the proposed regional relation module and cascaded pooling can effectively enhance their representative ability compared to directly optimizing the PA descriptors. We will show the performance of these three kinds of descriptors in Section 5.3.

5 EXPERIMENTS

In this section, we first introduce the experimental datasets, evaluation metric and implementation details. Then we give ablation study about the proposed relation features and cascaded pooling method. We also give qualitative analysis on the property of regional features and compare with state-of-the-art works on place recognition and image retrieval benchmarks. At last we combine relation features with NetVLAD aggregation method and provide comparative experiments.

5.1 Datasets

Our network architectures are trained on the Google Street View training datasets [2] and evaluated on the Pitts250k-test and Tokyo 24/7 dataset, respectively. All the images in the datasets above are of size $640 \times 480 \times 3$ except the queries of Tokyo 24/7.

Google Street View training datasets are composed of Pitts30k-train and Tokyo Time Machine datasets (Tokyo TM), which are sampled from Google Street View. Pitts30k-train contains 7k query images and 10k database images. Tokyo TM contains 7k query images and 49k database images sampled from different times in Tokyo area.

Pitts250k-test is a subset of the Pittsburgh dataset [56] and consists of around 83k database images and 8k query images generated from the panoramas in Pittsburgh area.

Tokyo 24/7 [53] has 76k database images sampled from the Google Street View and 315 query images taken from different mobile phone cameras in Tokyo area. Tokyo 24/7 dataset is very challenging because the streets are crowded in Tokyo area and the query images were taken at daytime, sunset and night while the database images were only taken at daytime. In Tokyo 24/7 dataset, there are less query images than the Pitts250k-test, so the performances on Tokyo 24/7 are usually change dramatically as shown in the following experiments.

Evaluation metric. For Pitts250k-test and Tokyo 24/7 dataset, we use Recall@N as evaluation metric, which is similar to the Rank-N accuracy in person re-identification [61]. A query image is deemed correctly recognized if at least one of the top N candidate images is within 25 meters from the ground truth GPS coordinates of the query. We calculate the mean Recall@N for all queries. All the images in the datasets above have spatial size of 640×480 except the queries of Tokyo 24/7 and when testing we resize the larger side of these queries to 640 while keeping the aspect ratio.

5.2 Implementation details

The pre-trained AlexNet [28] and VGGNet [48] are adopted as the base models which are both cropped at the last convolutional layer, before ReLU. We adopt pyramid pooling with 3-scale spatial grids (2×2, 4×4, 6×6), where the region number N in equation 4 is 56. For the MLP in relation module, we adopt a single convolutional layer to get relational feature maps. We separately choose $K = 3$ and 8 for Pitts250k and Tokyo 24/7 dataset because the scenes in Tokyo-24/7 dataset are more complex with more objects than Pitts250k-test. In the training process, we use margin m = 0.1, momentum 0.9, weight decay 0.001, batch size of 4 tuples, SGD with initial learning rate l_0 0.001 or 0.0005 for Pitts30k-train or Tokyo TM training set. The learning rate is equal to $l_0 \exp(-0.1(i-1))$, which is decayed exponentially over epoch i. Other hyper-parameters for training are the same as NetVLAD [2]. We perform PCA power-whitening for dimensionality reduction as in [63] and all the descriptors are ℓ_2-normalized for testing by default. All the descriptors for comparison in this section are end-to-end optimized unless otherwise specified.

5.3 Ablation study

5.3.1 Relational Feature Maps. To demonstrate the advantage of the relational feature maps ($\mathcal{R}$-FM), we adopt two commonly used aggregation methods (*i.e.* global max pooling (GMP) and global average pooling (sum pooling)) on $\mathcal{R}$-FM and convolutional feature maps (C-FM) to get the global descriptors for place recognition. As shown in the upper part of Table 1, $\mathcal{R}$-FM usually perform better than C-FM on both datasets, especially when sum pooling is adopted. The reason is that $\mathcal{R}$-FM contain high-level information about objects and their relationships, thus the relation descriptors are more discriminative. This demonstrates the benefits of our regional relation module and relation descriptors.

5.3.2 Cascaded pooling. In the lower part of Table 1, we also show the performance of our cascaded pooling method for aggregating

Table 1: Comparison of $\mathcal{C}$-FM and $\mathcal{R}$-FM when different aggregation methods are adopted. All these methods are trained end-to-end on the Pitts30k-train dataset. All these results are based on the VGGNet architecture. The best results are highlighted in bold.

Method	Feature maps	Pitts250k-test			Tokyo 24/7		
		Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
GMP	$\mathcal{C}$-FM	79.2	90.1	93.1	48.9	63.5	69.8
	$\mathcal{R}$-FM	81.7	91.3	93.8	47.6	64.4	74.0
Sum pooling	$\mathcal{C}$-FM	75.6	88.0	91.8	44.4	59.7	66.7
	$\mathcal{R}$-FM	82.0	91.3	94.0	55.9	72.4	75.6
Cascaded pooling ($K = 1$)	$\mathcal{R}$-FM	83.9	92.4	94.9	56.5	71.4	77.5
Cascaded pooling ($K = 3$)		**84.8**	**92.9**	**95.2**	61.9	77.8	83.2
Cascaded pooling ($K = 8$)		84.5	92.7	94.9	**67.3**	80.6	**85.1**
Cascaded pooling ($K = 15$)		83.2	92.0	94.4	65.7	**82.9**	84.8

Table 2: Comparison of different kinds of relation modeling methods on Pitts250k-test using AlexNet and VGGNet architectures. PANet is the baseline method which directly sum-aggregates the region feature set.

Method	AlexNet				VGGNet			
	Dim	R@1	R@5	R@10	*Dim*	R@1	R@5	R@10
PANet [63]	256	69.8	83.4	87.3	512	82.5	91.6	94.1
Similarity graph	512	70.2	84.5	88.5	1024	82.9	91.7	94.5
Spatial graph	512	70.4	84.3	88.0	1024	82.6	92.0	94.3
Ours	512	**77.2**	**87.9**	**90.7**	1024	**84.8**	**92.9**	**95.2**

the $\mathcal{R}$-FM. As we can see, cascaded pooling consistently outperforms two commonly used global max pooling (GMP) and sum pooling method on both datasets. The accuracy on Tokyo 24/7 dataset is largely increased as K in the cascaded pooling increases from 1 to 8, and drops lightly when $K = 15$. And the performance on Pitts250k-test is not so sensitive to K. It is because the streets are crowded in Tokyo 24/7 dataset and there are more objects in an image than Pitts250k-test, so sum pooling preserve more useful information than GMP and cascaded pooling further discards the noisy information in the relational region features. In the latter experiments, we choose $K = 3$ and $K = 8$ for Pitts250k-test and Tokyo 24/7 dataset, respectively.

5.3.3 Comparing with other relation modeling blocks. Recently, the Non-local block [57] and graph convolutional network (GCN) [25] are widely adopted for relation modeling in various vision tasks. We implement these two methods to model regional relation on Pitts250k-test dataset. Having the region feature set $\mathbf{f}_\Omega$ with shape of $N \times D$, we adopt a general version of the Non-local and graph convolutional function to model regional relation as follow:

$$\mathbf{f}_\Omega^{'} = G\mathbf{f}_\Omega W, \tag{8}$$

where G is the $N \times N$ graph representing the connection of N region features, W with shape of $D \times D$ is parameters to be optimized. For the graph G, we construct a similarity graph (corresponding to Non-local) and a spatial graph (corresponding to GCN) which separately represent the appearance relation and spatial relation of region features.

Similarity graph. The construction of similarity graph is similar to [57, 58]. Having the region features $\mathbf{f}_\Omega = \{f_{r,1}, f_{r,2}, ...f_{r,N}\}$ with shape of $N \times D$, the pairwise similarity between every two regions can be represented as

$$S(f_{r,i}, f_{r,j}) = \phi(f_{r,i})^T \phi^{'}(f_{r,j}), \tag{9}$$

where $\phi(x) = wx$ and $\phi^{'}(x) = w^{'}x$ represent two different transformations of the original features. In practice, they are separately two 1×1 convolutional layers with parameter shape of $D \times D$. All the pair-wise similarity $S(f_{r,i}, f_{r,j})$ construct the $N \times N$ affinity matrix, *i.e.* similarity graph. Each row of the affinity matrix is further normalized with softmax function so that the sum of all the edge values connected to one region i will be 1.

Spatial graph. The construction of spatial graph is similar to the spatial-temporal graph in [58]. Because we perform pyramid pooling to get the region features, the spatial coordinate of each region is known. The spatial coordinate (x, y, h, w) of a region denotes the central position in X-axis, Y-axis and the height, width of this region. Then we calculate the value of Intersection Over Unions (IoUs) of all region pairs and thus construct the $N \times N$ spatial graph. So the minimum value is 0 and the maximum value is 1 in the spatial graph. Each row of the spatial graph is further $L1$-normalized so that the sum of all the edge values connected to one region i will be 1.

After constructing the graph, we perform the graph convolution operation in equation 8 and get the relational region feature set $\mathbf{f}_\Omega^{'}$. In [57, 58], the graph convolution operation can be stacked several times, but we found there is performance drop when stacking more graph convolution layers. So we just perform the graph convolution operation once, then we fuse the original region features with relational region feature and sum-aggregate the fused region features to global descriptor. Specifically, we concatenate the region features with the relational region features and get C-dimensional ($C = 2 * D$) global relation descriptor after sum-aggregation.

We compare our relation descriptor with these two methods in Table 2. We can see that our method largely surpass the PANet baseline while the similarity and spatial graph convolution operations are not so effective to enhance the discriminative ability

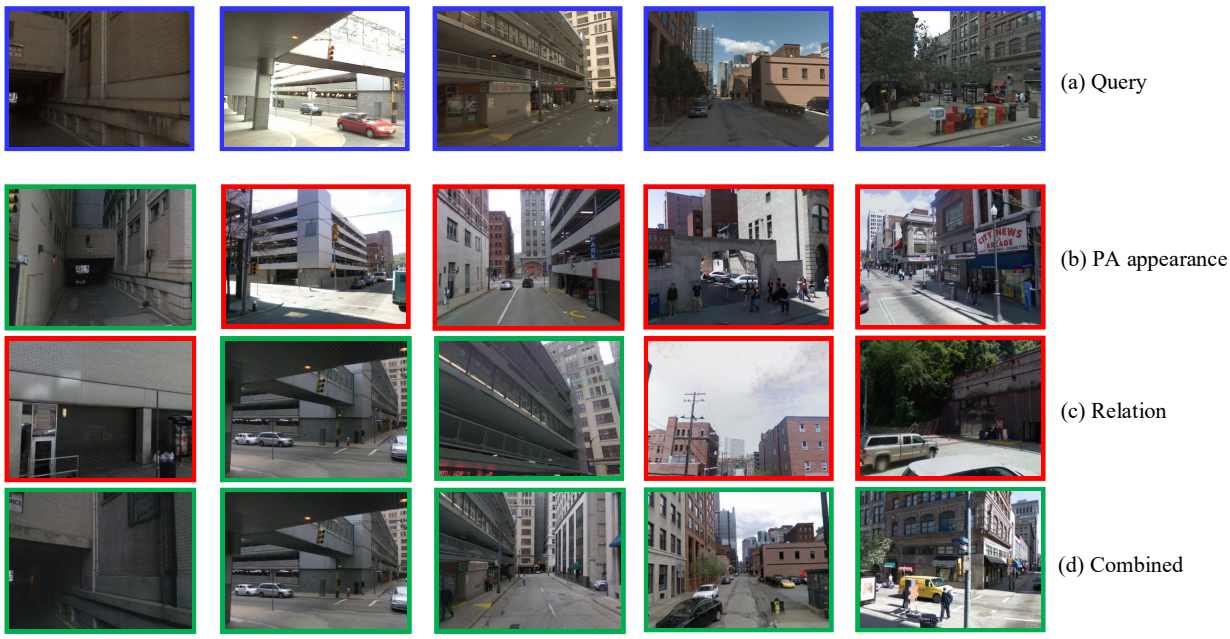


Figure 3: Retrieval examples. From top to bottom are the query images and Top-1 retrieval results of three kinds of descriptors, which are based on AlexNet because the performance gaps are larger than that of VGGNet. True matching is marked with green border while false matching is red.

of region features. We guess the reason is the receptive fields of region features are larger than the whole image and they are usually overlapped, the concatenation of region features and relation modeling in MLP can capture more statistical and dynamic relation than message passing in the GCN.

5.3.4 Combining relation with appearance. As described in Section 4.4, we choose the Pyramid Aggregation (PA) descriptors, which are generated by sum-aggregating the region features, as our appearance descriptors to form the combined descriptors with relation descriptors. We compare four kinds of descriptors on two datasets in Table 3 and Table 4, where "PANet" denotes the PA descriptors which are end-to-end optimized as in [63], and they can be considered as our baseline. "PA appearance" refers to the PA descriptors in our architecture, they are regularized by regional relation module and cascaded pooling and they form the "Combined" descriptors with "Relation" descriptors.

The result comparisons in two tables can be summarized into three observations. First, PA appearance descriptors outperform the end-to-end optimized PANet descriptors using two base CNN models, even they are just the outputs of an intermediate layer in our architecture. This observation is more distinct on Tokyo 24/7 dataset and demonstrates that the appearance descriptors are effectively improved when training with our regional relation module. Second, relation descriptors are superior to the PA appearance descriptors on Pitts250k-test dataset while are outperformed by PA appearance descriptors on Tokyo 24/7 dataset using AlexNet

Table 3: Comparison of different kinds of descriptors on Pitts250k-test dataset using AlexNet and VGGNet architectures. DR denotes dimension reduction.

Method	AlexNet				VGGNet			
	Dim	R@1	R@5	R@10	*Dim*	R@1	R@5	R@10
PANet [63]	256	69.8	83.4	87.3	512	82.5	91.6	94.1
PA appearance	256	70.5	83.8	87.6	512	82.6	92.2	94.6
Relation	512	77.2	87.9	90.7	1024	84.8	92.9	95.2
Combined	768	**79.3**	**89.4**	**92.2**	1536	**86.1**	**93.6**	**95.7**
Combined (DR)	256	75.7	87.9	90.8	512	84.6	93.2	95.4

Table 4: Comparison of different kinds of descriptors on Tokyo 24/7 dataset using AlexNet and VGGNet architectures. DR denotes dimension reduction.

Method	AlexNet				VGGNet			
	Dim	R@1	R@5	R@10	*Dim*	R@1	R@5	R@10
PANet [63]	256	36.5	51.1	57.5	512	56.5	68.3	74.3
PA appearance	256	39.4	**56.8**	**66.0**	512	65.1	81.0	86.7
Relation	512	37.5	52.1	57.1	1024	67.3	80.6	85.1
Combined	768	**47.0**	56.2	63.5	1536	**72.7**	**84.1**	**88.3**
Combined (DR)	256	40.3	52.7	62.2	512	68.3	82.2	87.6

architecture. We conjecture this is because there are more confusing objects (pedestrians and cars) on the Tokyo 24/7 dataset than

Table 5: Comparison with state-of-the-art methods. All these methods are based on VGGNet architecture. The best results are highlighted in red and best results in 512-d are highlighted in blue.

Method	*Dim*	Pitts250k-test			Tokyo 24/7			Tokyo 24/7 sunset/night		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
NetVLAD [2]	4096	86.0	93.2	95.1	71.8	82.5	86.4	61.4	75.7	81.0
CRN [24]	4096	85.5	93.5	95.5	**75.2**	83.8	87.3	**66.7**	76.7	81.9
NetVLAD [2]	1536	84.3	92.7	94.5	67.3	77.8	82.5	56.7	68.1	74.8
Our combined	1536	**86.1**	**93.6**	**95.7**	72.7	**84.1**	**88.3**	62.9	**77.6**	**83.3**
NetVLAD [2]	512	80.7	90.9	93.1	60.0	73.7	79.1	45.7	63.3	70.0
APANet [63]	512	83.7	92.6	94.7	67.0	81.0	83.8	**57.1**	**74.8**	78.6
Our combined	512	**84.6**	**93.2**	**95.4**	**68.3**	**82.2**	**87.6**	**57.1**	**74.8**	**82.4**

Pitts250k-test, and the appearance descriptors are usually more robust to the confusing objects. Even so, the relation descriptors still perform better than the baseline. Third, when concatenating the PA appearance and relation descriptors together, the combined descriptors show large performance improvements and still perform well after dimensionality reduction. It can be explained that appearance and relation descriptors focus on different cues in a place, so they complement each other for place recognition.

5.4 Qualitative analysis

In Figure 3, we show some representative retrieval results of the relation, PA appearance and combined descriptors. From these retrieval results, we have several observations to conclude. First, the PA appearance descriptors show translation invariance inherited from the convolution features and the orderless aggregation method. They show invariance to viewpoint changes (Column 1) but perform not so well on illumination or appearance changes (Column 2, 4). Second, the relation descriptors focus more on the spatial layout. The retrieval results of relation descriptors usually have less viewpoint changes (Column 3), and show invariance to the changes of appearance (Column 2). But the relation descriptors prone to be influenced by the confusing objects (Column 5). Third, combining the advantages of both descriptors, the combined descriptors take object appearance and object relations into consideration and get more reliable results.

5.5 Compared to state-of-the-art

Recently, NetVLAD-based deep descriptors [2, 24] have shown state-of-the-art performance on place recognition benchmarks and APANet descriptors [63] perform pretty well at low dimensionality. We compare our combined descriptors with these methods in Table 5. For Pitts250k-test dataset, our combined descriptors surpasses NetVLAD-based descriptors with approximately 3-times shorter descriptors. For Tokyo 24/7 dataset, our method is surpassed by CRN on Recall@1. The reason is CRN did additional data augmentation on the illumination conditions (two thirds of the queries in Tokyo 24/7 dataset were taken at sunset and night time) and three-clip testing, so it is unfair to compare CRN with other works. Even so, we still achieve best performance on Recall@5 and Recall@10. We can observe that for the lower dimensionality, the combined descriptors still outperform the NetVLAD and APnet descriptors on both datasets.

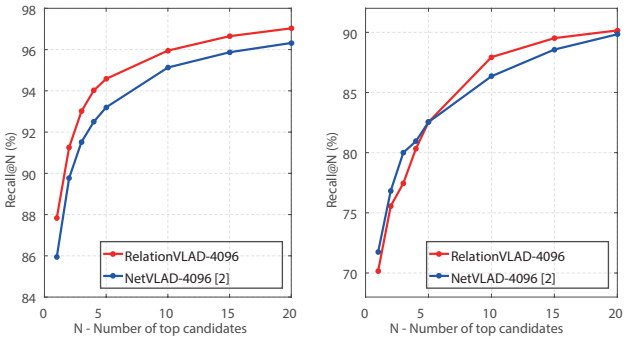
Note that our method has broad room for further improvements because it is compatible with other three competitors. We can adopt NetVLAD as a powerful aggregation method to aggregate the $\mathcal{R}$-FM, and we provide tentative experimental results in Section 5.7. We can also adopt the attention block of APANet or CRN to discover salient regions and region-pairs, and this is left for our future work.

Table 6: Comparisons of different methods on image retrieval datasets. The accuracy is measured by mean average precision (mAP) and all the results are on the basis of VGG-16 architecture, single-scale image descriptors.

Method	*Dim*	Oxford	Paris	Holidays
NetVLAD [2]	4096	71.6	79.7	87.5
CRN [24]	4096	69.2	-	-
APANet [63]	512	**77.9**	83.5	-
PA appearance	512	**77.9**	**84.9**	88.8
Relation	1024	75.0	81.3	86.3
Combined	1536	76.8	83.6	**88.9**

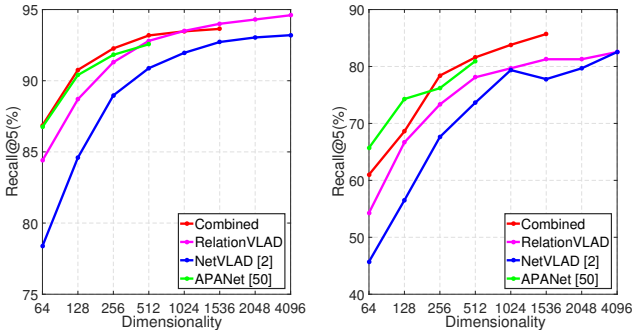
5.6 Image retrieval

To show the generalization ability of our method, we deploy the trained model (trained on Pitts30k-train dataset) on three standard image retrieval datasets: the Oxford 5k [38], Paris 6k [39] and Holidays [20], and show the results in Table 6. It is clear that our PA appearance and combined descriptors get best results on three datasets, and the relation descriptors are consistently exceeded by PA appearance descriptors on these datasets. The interpretation is that query images in the image retrieval datasets usually contain only one object, so the assumption of modeling object relation degrades to the relation modeling of object parts. Relation descriptors seem to be not so effective in these simple scenes, but they still perform better than the NetVLAD-based descriptors. We can also



(a) Pitts250k-test (b) Tokyo 24/7

Figure 4: Comparisons of NetVLAD and RelationVLAD descriptors. The dimensionality is followed.



(a) Pitts250k-test (b) Tokyo 24/7

Figure 5: Comparisons of NetVLAD-based and pooling-based descriptors about accuracy and dimensionality.

find our PA appearance descriptors exceed APANet descriptors, which additionally adopt attention block on the region features and are end-to-end optimized. This observation proves once again that training with our regional relation module effectively enhances the appearance descriptors. It is worth noting that on the Holidays dataset, PA appearance descriptors even surpass the state-of-the-art image retrieval works such as GeM (87.3) [40] and DIR (88.7) [15] under similar configurations (512-d VGGNet descriptors, without spatial re-ranking or query expansion). Here we don't include the results of the best image retrieval works [40, 42] because our model is trained on the street-view images and these works are trained on the landmark images, while the image retrieval datasets are landmark-centered. So we just fairly compare with the models trained on the street-view images to demonstrate the generalization ability of our method.

5.7 RelationVLAD

In this subsection, we provide tentative experiments that combine $\mathcal{R}$-FM with the VLAD [2, 21] aggregation method. Specifically, after converting the $\mathcal{C}$-FM to $\mathcal{R}$-FM by regional relation module, we adopt NetVLAD as a powerful aggregation method to get global relation descriptors and we call them "RelationVLAD". Note that end-to-end optimizing the RelationVLAD descriptors gets poor performance in our implementation. So we first optimize the relation descriptors as illustrated in Figure 2, then we replace the cascaded pooling module of the trained models with NetVLAD layer for aggregation. In this way we get relationVLAD descriptors with pretty good results, and we don't perform further optimization because this gets few improvements.

In figure 4 we compare RelationVLAD with NetVLAD descriptors which are generated by aggregating the local convolutional features. As can be seen, RelationVLAD descriptors consistently outperform NetVLAD descriptors on Pitts250k-test datasets but perform similarly with NetVLAD on Tokyo 24/7 dataset. We conjecture it is because relation features are prone to be influenced by the confused objects which are common in the streets of Tokyo area. It is worth mentioning that RelationVLAD descriptors usually perform better when the top database candidates N gets larger, which demonstrates the benefits of the relation features.

Since the NetVLAD-based and pooling-based aggregation methods get descriptors with different dimensions, we show the Recall@5 accuracies of these methods in various dimensions for fair comparison in Figure 5. It can be seen that the pooling-based methods usually perform better for both datasets in the same dimensions, especially in lower dimensions. On Pitts250k-test dataset, RelationVLAD descriptors perform best when the dimensionality is larger than 1024 but are surpassed by the pooling-based methods in lower dimensions. Our combined descriptors perform better than APANet on Pitts250k-test dataset but are exceeded on Tokyo 24/7 in lower dimensions.

6 CONCLUSION AND FUTURE WORK

In this paper, we consider the visual relationships as important cues for visual place recognition. We propose a regional relation module which imitates the relation modeling process of visual relationship modeling framework and converts the $\mathcal{C}$-FM to $\mathcal{R}$-FM. The $\mathcal{R}$-FM contain rich information about objects and object relations. We further design a cascaded pooling method to effectively aggregate the $\mathcal{R}$-FM. Experiment evaluations demonstrate the effectiveness of our contribution and the properties of relation features. Potential improvements could include the use of attentional mechanisms to explore salient regions and regional pairs, as well as generalizing our approach to single-object image retrieval dataset. Furthermore, our method has extensive connections with the visual relation detection, visual question answering, self-supervised learning tasks and we will reference their successful efforts in the future works.

ACKNOWLEDGMENTS

This work was supported by: (i) National Natural Science Foundation of China (Grant No. 61602314); (ii) Natural Science Foundation of Guangdong Province of China (Grant No. 2016A030313043); (iii) the Major Fundamental Research Project in the Science and Technology Plan of Shenzhen (Grant No. JCYJ20190808172007500).

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*. 2425–2433.

[2] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*. 5297–5307.

[3] Relja Arandjelović and Andrew Zisserman. 2013. All About VLAD. In *CVPR*. 1578–1585.

[4] Relja Arandjelović and Andrew Zisserman. 2014. DisLocation: Scalable descriptor distinctiveness for location recognition. In *ACCV*. 188–204.

[5] Artem Babenko and Victor Lempitsky. 2015. Aggregating local deep features for image retrieval. In *ICCV*. 1269–1277.

[6] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. 2016. Interaction networks for learning about objects, relations and physics. In *NIPS*. 4502–4510.

[7] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *ECCV*. 404–417.

[8] David M. Chen, Georges Baatz, Kevin Koser, Sam S. Tsai, and Radek Grzeszczuk. 2011. City-scale landmark identification on mobile devices. In *CVPR*. 737–744.

[9] Zetao Chen, Fabiola Maffra, Inkyu Sa, and Margarita Chli. 2017. Only look once, mining distinctive landmarks from ConvNet for visual place recognition. In *IEEE/IROS*.

[10] Mark Cummins and Paul Newman. 2008. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *IJRR* 27, 6 (2008), 647–665.

[11] Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. In *CVPR*. 3298–3308.

[12] Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *ICCV*. 1422–1430.

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *IJCV* 88, 2 (2010), 303–338.

[14] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. 2016. Deep image retrieval: Learning global representations for image search. In *ECCV*. 241–257.

[15] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. 2017. End-to-end learning of deep visual representations for image retrieval. *IJCV* 124, 2 (2017), 237–254.

[16] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. *arXiv preprint arXiv:1802.08218* (2018).

[17] James Hays and Alexei A Efros. 2008. IM2GPS: estimating geographic information from a single image. In *CVPR*. 1–8.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[19] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. 2018. Relation Networks for Object Detection. In *CVPR*. 3588–3597.

[20] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2008. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*. 304–317.

[21] Herve Jegou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Perez, and Cordelia Schmid. 2012. Aggregating local image descriptors into compact codes. *TPAMI* 34, 9 (2012), 1704–1716.

[22] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. 2015. Predicting good features for image geo-localization using per-bundle vlad. In *ICCV*. 1170–1178.

[23] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*. 1988–1997.

[24] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. 2017. Learned contextual feature reweighting for image geo-localization. In *CVPR*. 2136–2145.

[25] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[26] Jan Knopp, Josef Sivic, and Tomas Pajdla. 2010. Avoiding confusing features in place recognition. *ECCV*, 748–761.

[27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S Bernstein, and Fei-Fei Li. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *IJCV* 123, 1 (2017), 32–73.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*. 1097–1105.

[29] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated graph sequence neural networks. In *ICLR*.

[30] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. 2019. VrR-VG: Refocusing Visually-Relevant Relationships. In *Proceedings of the IEEE International Conference on Computer Vision*. 10403–10412.

[31] Liu Liu, Hongdong Li, and Yuchao Dai. 2017. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *ICCV*. 2391–2400.

[32] Yihang Lou, Yan Bai, Shiqi Wang, and Ling-Yu Duan. 2018. Multi-Scale Context Attention Network for Image Retrieval. In *ACM MM*. 1128–1136.

[33] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *IJCV* 60, 2 (2004), 91–110.

[34] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *ECCV*. 852–869.

[35] Colin McManus, Winston Churchill, Will Maddern, Alexander D Stewart, and Paul Newman. 2014. Shady dealings: Robust, long-term visual localisation using illumination invariance. In *ICRA*. 901–906.

[36] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. 2017. Largescale image retrieval with attentive deep local features. In *ICCV*. 3456–3465.

[37] Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*. 69–84.

[38] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*. 1–8.

[39] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*. 1–8.

[40] Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2018. Fine-tuning CNN Image Retrieval with No Human Annotation. *TPAMI* (2018).

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*. 91–99.

[42] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. 2019. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE International Conference on Computer Vision*. 5107–5116.

[43] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. 2017. A simple neural network module for relational reasoning. In *NIPS*. 4967–4976.

[44] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. 2017. Efficient & effective prioritized matching for large-scale image-based localization. *TPAMI* 9 (2017), 1744–1756.

[45] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*. 815–823.

[46] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. 2018. CPlaNet: Enhancing Image Geolocalization by Combinatorial Partitioning of Maps. In *ECCV*. 544–560.

[47] Oriane Siméoni, Yannis Avrithis, and Ondrej Chum. 2019. Local Features and Visual Words Emerge in Activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11651–11660.

[48] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

[49] Josef Sivic and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *ICCV*. 1470–1477.

[50] Abby Stylianou, Richard Souvenir, and Robert Pless. 2019. Visualizing Deep Similarity Networks. In *WACV*.

[51] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. 2016. Image search with selective match kernels: aggregation across single and multiple images. *IJCV* 116, 3 (2016), 247–261.

[52] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2016. Particular object retrieval with integral max-pooling of CNN activations. In *ICLR*.

[53] Akihiko Torii, Relja Arandjelović, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 2015. 24/7 place recognition by view synthesis. In *CVPR*. 1808–1817.

[54] A Torii, R Arandjelovic, J Sivic, M Okutomi, and T Pajdla. 2018. 24/7 Place Recognition by View Synthesis. *TPAMI* 40, 2 (2018), 257.

[55] Akihiko Torii, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 2015. Visual Place Recognition with Repetitive Structures. *TPAMI* 37, 11 (2015), 2346–2359.

[56] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. 2013. iVsual place recognition with repetitive structures. In *CVPR*. 883–890.

[57] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7794–7803.

[58] Xiaolong Wang and Abhinav Gupta. 2018. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*. 399–417.

[59] Xu Yang, Hanwang Zhang, and Jianfei Cai. 2018. Shuffle-Then-Assemble: Learning Object-Agnostic Visual Relationship Features. In *ECCV*. 38–54.

[60] Peng Yin, Lingyun Xu, Xueqian Li, Chen Yin, Yingli Li, Rangaprasad Arun Srivatsan, Lu Li, Jianmin Ji, and Yuqing He. 2019. A Multi-Domain Feature Learning Method for Visual Place Recognition. *arXiv preprint arXiv:1902.10058* (2019).

[61] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *ICCV*. 1116–1124.

[62] Liang Zheng, Yi Yang, and Qi Tian. 2017. SIFT meets CNN: A decade survey of instance retrieval. *TPAMI* (2017).

[63] Yingying Zhu, Jiong Wang, Lingxi Xie, and Liang Zheng. 2018. Attention-based Pyramid Aggregation Network for Visual Place Recognition. In *ACM MM*. 99–107.